On the Impossibility of Localizing Multiple Rumor Sources in a Line Graph

Sam Spencer
Coordinated Science Lab, Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
samster@illinois.edu

R. Srikant
Coordinated Science Lab, Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
rsrikant@illinois.edu

ABSTRACT

Here we examine the problem of rumor source identification in line graphs. We assume the SI model for rumor propagation with exponential waiting times. We consider the case where a rumor originates from two sources simultaneously, and evaluate the likelihood function for the given observations given those sources. As the size of the infected region grows arbitrarily large, we show that unlike the single source case, where the likelihood function concentrates near the midpoint of the infected region, the support of the likelihood function in this case remains widely distributed over the middle half of the infected region. This makes the rumor sources impossible to localize with high probability on any scale smaller than that of the infection size itself.

Categories and Subject Descriptors

G.2.2 [**Graph Theory**]: Network problems, Graph algorithms; G.2.1 [**Combinatorics**]: Combinatorial algorithms, Counting problems, Permutations and combinations.

General Terms

Algorithms, Performance, Theory, Security.

Keywords

Epidemics, Estimation.

1. INTRODUCTION

Rumor propagation and source detection problems (and their mathematical analogs) arise in a variety of contexts, including cybersecurity, information assurance, privacy, epidemiology, and social network analysis. Here we consider the problem of rumor source detection in undirected line graphs. Given a snapshot of the state of the network at an unknown point in time, we wish to estimate the likelihood function of that state ever arising from a given set of sources. For a uniform prior, this is equivalent to estimating the probability that these were the true sources.

Similar problems and approaches have been considered previously. In [2], the authors present the problem of maximum likelihood estimation of a single source of a rumor which propagates according to the SI model. Despite its structural simplicity, a line graph is actually shown to be a tougher challenge than trees of higher degree. Here, we will generalize the approach to consider a rumor with two sources. In [3], the authors consider the problem of multiple sources in a tree under the SIR model, and where the infection process occurs in discrete time. While the use of the SIR model actually allows for more generality in the infection process, their method does not apply to trees of degree 2, which is equivalent to our line graph model. Other current work in this area includes [1], which looks at this problem from the opposite side — trying to conceal the source of the rumor by manipulating the propagation model.

Research supported in part by AFOSR MURI FA 9550-10-1-0573.

2. MODEL

Our propagation model is as follows. We represent our line graph (a regular tree of degree 2) as the set of integers on the number line. We use the SI infection model with edge-based propagation in continuous time to describe the spreading of the rumor. That is, nodes are either "susceptible" (have not yet heard the rumor) or "infected" (have already heard it). If a susceptible node shares an edge with an infected neighbor, then the infection will "traverse" that edge and infect the susceptible node with a waiting time that is exponentially distributed with mean T. This means that if a susceptible node has two infected neighbors, then the waiting time for it to become infected remains exponential, but the mean drops to $T/2$. Once infected, a node remains that way indefinitely. An important consequence of this model is that we can invoke the memoryless property of the system to state that at any given time, the next infection to occur is equally likely along any outgoing edge from the current infected set.

2.1 Initial Setup

Our initial infections take place at two nodes simultaneously, at an unknown time. If the initially infected nodes are not adjacent to each other, subsequent infections can proceed from either source (or an interval of infected nodes containing that source), and spread either to the left or the right, until the last uninfected node between the two intervals becomes infected, at which point subsequent infections can proceed from either end of the unified interval. If the initially infected nodes are adjacent, then infections proceed according to the "unified interval" case from the outset. We assume that at our time of interest, the infected region is a single interval of integers on the number line. If instead the infection were still two distinct intervals, we could be

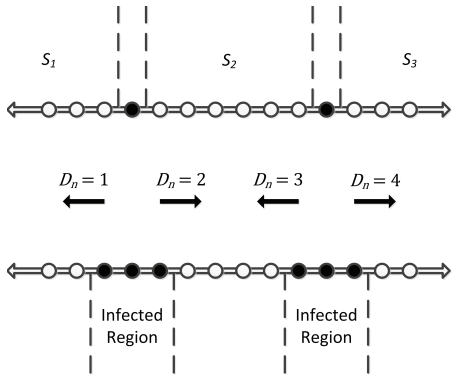

Figure 1: Infection sequence possibilities and uninfected regions. The values of D_n correspond to the four possible ways the infection can grow.

sure that each interval had its own (single) source, and simply compute the likelihood function for the source of each individual interval using the approach in [1], resulting in a binomial distribution over the nodes of each interval. It should be noted that while the probability of correctly identifying the exact source node within such an interval goes to 0 as the size of the infection increases (assuming a uniform prior), the source node can be localized with high probability within a region whose size goes to zero as a fraction of the size of the infection. We will further assume that our single infected interval is of even size. For simplicity, we will represent our eventual infected set as the interval $[1, 2K+2]$ for some value of K. That way, once we place the two initial infections, there are $2K$ remaining infections that need to occur in order to reach our observed end state.

2.2 Sequential Evolution

Next, rather than considering the evolution of the system in a temporal sense per se, we will invoke the memoryless property of the system, and consider only the sequence in which infections occur. Beginning from the initial two nodes, we will define an i.i.d. sequence of random variables D_n for $n = 1, 2, 3, \ldots$, which take on the values $\{1, 2, 3, 4\}$ with equal probabilities. A value of 1 corresponds to the next infection proceeding to the left from the leftmost origin or interval (specifically, from the leftmost infected point of the left interval), 2 corresponds to a rightward infection from the left interval, 3 corresponds to a leftward infection from the right interval, and 4 corresponds to a rightward infection from the right interval. This is shown in Figure 1.

Note that once the two intervals become joined, it is no longer possible for the infections corresponding to 2 and 3 to occur. However, for our purposes, these cases will make no difference, so we will simply state for now that such extraneous values of D_n will have no physical interpretation.

3. ANALYSIS

It will also be useful for us to define a set of counting functions $C_1(N)$, $C_2(N)$, $C_3(N)$ and $C_4(N)$, which simply count the number of instances in which D_n takes on the corresponding values as n ranges from 1 to N. Formally, let $C_l(N) = \sum_{n=1}^{N} \mathbb{1}_{\{D_n = l\}}$. Note that these functions are monotonically increasing in N, and thus so are any sums of them.

3.1 Success Conditions

By using these functions, we can make precise the necessary and sufficient conditions needed for infections at the two initial sources to lead to our desired end state at some point, without requiring an explicit dependence on time. For the moment, let S_1 be the set of uninfected nodes in $[1, 2K+2]$ to the left of the leftmost source, S_2 be the set of uninfected nodes between the sources, and S_3 be the set of uninfected nodes to the right of the rightmost source, as shown in Fig. 1. Since we dont know the time at which our end state would be reached, we must consider intermediate states, and evaluate their interpretation with respect to our desired end state. Since we know that once a node is infected, it remains that way, then if any node which is uninfected in our desired end state is infected in an intermediate state, then that state cannot grow into the desired state. Accordingly, our infection must reach every node we desire to be infected before reaching any node which we do not. If so, then our desired end state has been reached, and if not, then it has not, and can never be. Since the sets of presently uninfected nodes that we desire to see infected are S_1, S_2, and S_3, and the first undesired nodes that would be infected are those immediately to the left of S_1 or immediately to the right of S_3, then we can decompose the above necessary and sufficient requirement for reaching the desired end state into two conditions, which are:

Condition 1. The two end intervals S_1 and S_3 must each become completely infected before the other end exceeds its bound. This is equivalent to saying that when the total number of infections in these two regions reaches their total size, the number of infections in each must match its size exactly. Formally, let N_0 be the last time at which $C_1(N_0) + C_4(N_0) = |S_1| + |S_3|$. Then we require that $C_1(N_0) = |S_1|$ and $C_4(N_0) = |S_3|$.

Condition 2. The middle interval S_2 must be completely filled before either of the end intervals exceeds its bound. Let N_1 be the first time that $C_2(N_1) + C_3(N_1) = |S_2|$. If we assume Condition 1 to be true, then the first time one of the ends will exceed its bound is at $N_0 + 1$. Therefore, we require $N_1 < N_0 + 1$.

Since the second condition assumes the first, the probability of both conditions being satisfied is simply the product of the two individual probabilities. Note that $|S_1|+|S_2|+|S_3| = 2K$. Given Condition 1, it should be intuitively clear that the most likely cases will be those where $|S_1| = |S_3|$. Since the probabilities that D_n takes on 1 or 4 are equal, then for a given total sum $|S_1| + |S_3|$, the allocation of that sum between $C_1(N_0)$ and $C_4(N_0)$ is a binomial distribution, and the most likely partitioning of values is for $C_1(N_0) = C_4(N_0)$. (assuming an even total sum). This maximizes the probability of satisfying Condition 1, and since we have not changed $|S_2|$, the probability of satisfying Condition 2 has not changed either. Thus, we have maximized the overall probability of success by positioning the sources equidistant from the two ends of the infected region. (The case where $|S_1| \neq |S_3|$ is not shown here, but the corresponding probability rolls off as $e^{-\frac{\alpha(|S_1|-|S_3|)^2}{2K}}$ for small differences.)

Therefore, we will hereafter let $|S_1| = |S_3| = k_1$, $|S_2| =$

$2k_2$, and $k_1 + k_2 = K$. Going forward, we are interested in the optimal partitioning of K between k_1 and k_2.

3.2 Results

We now compute the probability of satisfying Condition 1. To do so, let us consider only the values of n for which $D_n = 1$ or 4, and find the probability that $C_1(N_0) = C_4(N_0) = k_1$ given that $C_1(N_0) + C_4(N_0) = 2k_1$. This is equivalent to the probability, given $2k_1$ objects that are independent and equally likely to be of one kind or another, that exactly k_1 are of one kind and the remaining k_1 are of the other. Since there are $\binom{2k_1}{k_1}$ ways to choose the k_1 objects of a single kind, and a total of 2^{2k_1} assignments of the objects to the two classes, the probability P_1 of satisfying Condition 1 is

$$P_1(k_1) = \frac{\binom{2k_1}{k_1}}{2^{2k_1}} = \frac{(2k_1)!}{k_1!k_1!2^{2k_1}}.$$

Using Stirling's approximation, $n! = \sqrt{2\pi n}(\frac{n}{e})^n$, this can be approximated by $\frac{1}{\sqrt{\pi k_1}}$.

For Condition 2, we can aggregate the outcomes $D_n = 1$ and 4 together, and $D_n = 2$ and 3 together. We then wish to know the probability that $C_2(N) + C_3(N) = 2k_2$ before $C_1(N) + C_4(N) = 2k_1 + 1$. While at first glance this could appear to be difficult due to the uncertain end time, we can simply consider the case where $N = 2k_1 + 2k_2 = 2K$. Since $(C_1(2K)+C_4(2K))+(C_2(2K)+C_3(2K)) = 2K$, then either $C_2(2K)+C_3(2K) \geq 2k_2$ (in which case $C_1(2K)+C_4(2K) < 2k_1 + 1$), or $C_1(2K) + C_4(2K) \geq 2k_1 + 1$ (in which case $C_2(2K) + C_3(2K) < 2k_2$). Thus, while the end state may actually be reached earlier, it can be uniquely determined by observing the propagation sequence at precisely this stage. Since each of these two aggregated outcomes is equally likely, and $k_1 + k_2 = K$, we can consider the probability that given 2K independent objects equally likely to be of one kind or another, that no more than $2k_1$ are of a specified kind. This can be written in terms of the regularized incomplete beta function as $F(2k_1; 2K, 0.5)$. While space does not permit a derivation here, it can be shown that

$$\lim_{K\to\infty} F(2k_1; 2K, 0.5) = \begin{cases} 0 & \text{if } k_1 < 0.25(2K); \\ 0.5 & \text{if } k_1 = 0.25(2K); \\ 1 & \text{if } k_1 > 0.25(2K). \end{cases}$$

Accordingly, we can multiply the probabilities for Condition 1 and Condition 2, and conclude that as $K \to \infty$, the total probability of observing the desired pattern of a single interval of $2K$ infected nodes at any point given two simultaneous original infections at k_1 spots from each end can be written asymptotically as

$$\lim_{K\to\infty} P(k_1, K) = \begin{cases} 0 & \text{if } k_1 < 0.25(2K); \\ \frac{0.5}{\sqrt{\pi k_1}} & \text{if } k_1 = 0.25(2K); \\ \frac{1}{\sqrt{\pi k_1}} & \text{if } k_1 > 0.25(2K). \end{cases}$$

4. SIMULATION

The foregoing model was tested by constructing a simulation model in MATLAB. Results are shown in Figure 2 for instances of $2K$ ranging from 62 to 50,000. Note the increasingly steep drop-off on the left side of $k_1/2K = 0.25$, but the steady, gradual decay to the right. These empirical results agree with the findings in the previous section.

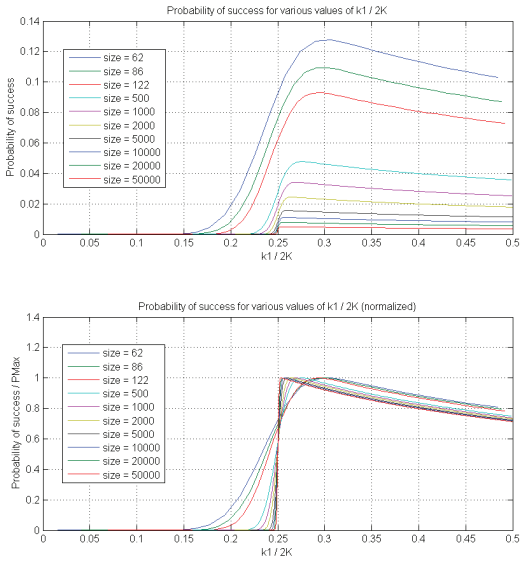

Figure 2: Simulation results validating theoretical findings. The top graph shows the results for various infection sizes, while the lower graph shows results normalized by the maximum value for each infection size.

5. COMPARISON AND CONCLUSION

For a single source, it can be shown using the approach from [1] that the likelihood function for the source given the observation follows a binomial distribution, which converges to a Gaussian whose standard deviation grows as $\sqrt{K}$. This means that, as K increases, the source can be localized with high probability within a region whose size grows as $\sqrt{K}$, which becomes vanishingly narrow relative to large values of K. In contrast, for the two source case, we have shown that for a given desired probability of correctness, the size of the region within which the sources can be located grows as K, and thus takes on an relatively constant fraction of the entire range (specifically, the middle half). This makes localization of the source essentially impossible.

6. REFERENCES

[1] G. Fanti, P. Kairouz, S. Oh, and P. Viswanath. Spy vs. spy: Rumor source obfuscation. In *Proceedings of the 15th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '15, to appear, 2015.

[2] D. Shah and T. Zaman. Rumor centrality: A universal source detector. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, pages 199–210, New York, NY, USA, 2012. ACM.

[3] K. Zhu and L. Ying. Information source detection in the sir model: A sample path based approach. In *Information Theory and Applications Workshop (ITA), 2013*, pages 1–9, Feb 2013.